Team Formation in Partially Observable Multi-Agent Systems

Adrian K. Agogino
University California, Santa Cruz
NASA Ames Research Center
Mailstop 269-4
Moffett Field, CA 94035
E-mail: adrian@email.arc.nasa.gov

Kagan Tumer
Computational Sciences Division
NASA Ames Research Center
Mailstop 269-4
Moffett Field, CA 94035
E-mail: ktumer@mail.arc.nasa.gov

Abstract— Sets of multi-agent teams often need to maximize a global utility rating the performance of the entire system where a team cannot fully observe other teams' agents. Such limited observability hinders team-members trying to pursue their team utilities to take actions that also help maximize the global utility. In this article, we show how team utilities can be used in partially observable systems. Furthermore, we show how team sizes can be manipulated to provide the best compromise between having easy to learn team utilities and having them aligned with the global utility. The results show that optimally sized teams in a partially observable environments outperform one team in a fully observable environment, by up to 30%.

I. INTRODUCTION

Team formation is important in many multi-agent systems, since it allows team members to focus on a team-goal, which is simpler than a global-goal over an entire system[4], [9]. In addition team formation allows sharing of information [6], [12]. This paper focuses on systems using teams of learning agents with the following properties:

- an agent belongs to one and only one team;
- agents receive the utility of the team; and
- team members share observations about other teams.

In this system each agent attempts to maximize a utility provided by the team using a learning algorithm such as reinforcement learning or evolution over neural networks. For such a system to work properly, team utilities have to have the following properties:

- team utilities should be easy for the agents to optimize;
- agents optimizing their team utilities should result in agents optimizing the global utility; and
- teams must compute utilities when they cannot fully observe each other.

In this paper we address the first property by using the theory of collectives [20], [15] to create learnable team utilities. We address the second and third property by modifying the theory of collectives for partially observable environments to create team utilities that are "aligned" with the global utility. We will show that these properties are traded off as the team size grows. The team utility for a small team is relatively easy to learn. However a large team is able to observe the actions of other teams better, allowing its team utility to be more aligned

with the global utility. We will show that team sizes can be adjusted to make the optimal tradeoff.

The concept of teams can be found most often in human activities. For example, corporations are often setup with team structures where employees are members of a team or group (e.g., through sharing a bonus for successful completition of a project) and each team member benefits when the team successfully contributes to the goals of the corporation. Spontaneous team formation in agents has also been studied at a theoretical level. Axtell [2] has shown that for small sizes of teams there can be a stable Nash equilibrium, but that the stability breaks down when teams go beyond a certain size. Similarly we will show that even when team formation is created in a top-down manner that it may be difficult for agents to learn to maximize team utilities for larger teams.

There has been extensive research on rule-based agent team formations. Tambe has shown that coordination rules can be used successfully in many fields including military engagement [14]. A common mechanism to coordinate team agents is for teams to have "joint intentions" [4] where team agents need to work for a common goal. Groz coins the term "SharedPlan" [9] to refer to this concept. Also related to this paper is work done in the field of sensor fusion. Fox has shown that when the amount of information that a robot receives is restricted teams of robots with different sensors, can work together to solve the robot localization problem [6]. In addition it has been shown that teams can share sensor information to estimate unobservable parts of the world in robotic soccer domains [12].

The first step in creating a collection of teams that can effectively maximize the global utility is to ensure that teams can work together. If the teams are not designed to work well with each other, they may not learn their task properly, may interfere with each other's ability to contribute to the global utility, or simply perform useless repetitive work. Hand tailoring the team utility functions may offer an alternative, but such systems: (i) have to be laboriously modeled; (ii) provide "brittle" global performance; (iii) are not "adaptive" to changing environments; and (iv) generally do not scale well.

To sidestep these problems, yet address the design requirements listed above (i.e., "alignedness" and "learnability") one

can use the framework of collectives [17], [20]. Given this framework, the crucial design problem becomes: Assuming the individual agents are able to maximize the team utility function (e.g., through reinforcement learning [13] or evolution of neural networks), what set of team utilities, when pursued by those agents, result in high global utility?

There are two quantifiable properties (discussed in detail in Section II) that help answer this question. First, the utility functions for the team need to be "aligned" with the global utility, in that an action taken by an agent that improves its team utility also improves the global utility. Second, the utility functions need to be "learnable" in that an agent has to be able to discern the effect of its actions on its team utility and select actions that optimize that utility. As we will highlight below, the theory of collectives provides utilities for agents that maximize the second property while satisfying the first one.

The collectives framework has been successfully applied to multiple domains including packet routing over a data network [21], the congestion game known as Arthur's El Farol Bar problem [22], and the coordination of multi-rovers in learning sequences of actions [16]. In particular, in the routing domain, the collectives approach achieved performance improvements of a factor of three over the conventional Shortest Path Algorithm (SPA) routing algorithms currently running on the internet [19], and avoided the Braess' routing paradox which plagues the SPA-based systems [17].

In the work described above, agents can fully observe each other and did not form teams. In this paper we will show that teams can be effective in environments with partial observability, if the proper team utilities are used. In Section II, we provide some background on the theory of collectives that is needed for this article. In Section III, we describe the problem domain and present the collective-based solution to this problem. In Section IV, we present the simulation results for domains where there is no costs associated with team members sharing informations and domains where there is a cost of sharing information.

II. BACKGROUND: COLLECTIVE INTELLIGENCE

In this work, we focus on a system of multi-agent teams that aim to maximize a global utility function, $G(z)$, which is a function of the joint move of all agents in the system, z. In previous work [20] that uses the theory of collectives, each agent does not maximize $G(z)$ directly, but instead maximizes an agent specific utility function, $g(z)$. Instead in this work each agent in team τ will try to maximize a **team utility function** $g_\tau(z)$. Team utilities have the advantage over agent utilities in partially observable environments in that a team utility may be able to incorporate observations from all the team members. This increase in observational capability will allow team utilities that are more "aligned" with the global utility (Figure 1). In addition team utilities allow for domains where agents are not even capable of computing their own utility, but can still blindly maximize a broadcast team utility. The goal of this section is to create team utility functions that

will cause the multi-agent system to produce high values of $G(z)$.
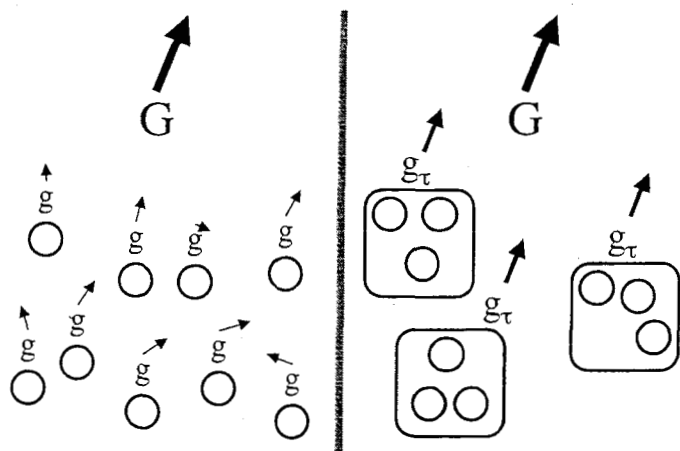

Fig. 1. Team vs. Agent Utilities. Left figure shows agents following agent utilities that are not fully aligned with the global utility due to partial observability. Right figure shows teams collecting observations from multiple agents allowing them to make team utilities that are more aligned.

Note that in many systems, an individual team τ will only influence some of the components of z. We will use the notation z_τ to refer to the parts of z that are dependent on the actions of team τ. The vector z_τ is the same size as z and is equal to z except that all the components that do not depend on team τ are set to zero. By subtracting z_τ from z we produce the vector $z_{-\tau} = z - z_\tau$, a vector that is determined by the actions of all the agents other than τ. Note that this subscripted vector notation is not the same as a traditional index to a vector since z, $z_{-\tau}$ and z_τ all have the same number of components.

There are two properties that are crucial to producing systems in which agents acting to optimize their team utilities will also optimize the provided global utility. The first of these concerns "aligning" the team utilities with the global utility. Formally, a system is **factored** when for each team τ:

$$g_\tau(z) \geq g_\tau(z') \Leftrightarrow G(z) \geq G(z')$$
$$\forall z, z' \ s.t. \ z - z_\tau = z' - z'_\tau \ .$$

Intuitively, for all pairs of states z and z' that differ only for team τ, a change in τ's state that increases its team utility cannot decrease the global utility. As a trivial example, any system in which all the team utility functions equal G is factored [5].

The second property, called **learnability**, measures the dependence of a utility on the actions of agents in a particular team as opposed to all the actions of all the other agents. Intuitively, higher learnability means it is easier for a team τ to achieve a large values of its team utility. Note in the factored example of using G as team utility above, the utility of each team depended on the actions of all the agents in all teams. Such systems often suffer from low signal-to-noise, a problem that get progressively worse as the size of the system grows.

A. Difference Utility

Consider **difference** utility functions, which are of the form:

$$DU_\tau \equiv G(z) - G(z_{-\tau} + c_\tau) \qquad (1)$$

where $z_{-\tau}$ contains all the variable not affected by agents in team τ. All the components of z that are affected by agents in team τ are replaced with the fixed constant c_τ. Such difference utilities are factored no matter what the choice of c_τ, because the second term does not depend on the actions of agents in team τ [20]. Furthermore, they usually have far better learnability than does the global utility, because of the second term of DU, which removes a lot of the effect of other agents (i.e., noise) from τ's evaluation function. In many situations it is possible to use a c_τ that is equivalent to taking team τ out of the system. Intuitively this causes the second term of the difference utility to evaluate the global utility of the system without team τ and therefore DU evaluates the teams contribution to the global utility.

B. Partial Observability

In general to compute a difference utility, a team may have to be able to fully observer all the other teams. For some specific classes of utility such as the DU, this observational demand may be relaxed, since many of the elements of the worldline cancel out and may be ignored. However in many real world problems, agents from one team cannot observe agents from other teams adequately to compute even the less demanding utilities. In these cases we must approximate the utility under the constraints of partial observability.

We denote the component of z that is observable by τ using the vector z^{o_τ} and the part of z that is not observable by τ using the vector z^{h_τ}. The vector z^{o_τ} is the same as z except that all the elements that are not observable by τ are set to zero. We call z^{o_τ} the **observable components** of the worldline. The vector z is the sum of these two vectors: $z = z^{o_\tau} + z^{h_\tau}$. It is assumed that team τ can always observe all components of z_τ. If the DU depends on any component of z^h then we cannot compute it directly. Instead there are several approximations to the DU that vary in their balance between learnability and factoredness. In this paper we discuss four approximations [1]:

$$BTU_\tau(z) = G(z) - G(z^{o_\tau} - z_\tau) \qquad (2)$$

$$TTU_\tau(z) = G(z^{o_\tau}) - G(z^{o_\tau} - z_\tau) \qquad (3)$$

$$BEU_\tau(z) = G(z) - G(z^{o_\tau} + E[z^{h_\tau}|z^{o_\tau}] - z_\tau) \qquad (4)$$

$$EEU_\tau(z) = G(z^{o_\tau} - E[z^{h_\tau}|z^{o_\tau}]) -$$
$$G(z^{o_\tau} + E[z^{h_\tau}|z^{o_\tau}] - z_\tau) , \qquad (5)$$

where $E[\cdot]$ is the expectation operator. Note that BTU, as well as BEU, assume that the true global utility can be broadcast despite having only partial observability. In many

[1]The first two letters of the utility represent how the two terms of the difference utility get their information. "B" stands for "broadcast", "T" ' stands for "truncated" since the hidden values are just thrown away, and "E" stands for "estimated."

applications, this is a reasonable assumption since the global utility can often be computed once and broadcast throughout the environment [7]. More complex forms of broadcasting are often used for distributed multi-agent systems [3], but in this paper we will assume a very simple global broadcast of a single number. In many domains it is also reasonable to assume global utility can even be obtained directly from the environment without broadcasting [10].

C. Observability and Team Size

This paper assumes that the observational capability of a team goes up with the size of the team. This property of team size can happen for a number of different reasons including, larger teams having more resources and greater coverage of different areas. This paper will use a simple model of how the observation capability of a team relates to its size, based on all team observations coming independently from team members. Let S_i be the set of components in z that the ith agent in team τ can observe. We define the agent observation level B_i as the ratio of the size of S_i to the size of z:

$$B_i = \frac{|S_i|}{|z|} . \qquad (6)$$

Note that this value is always in the range $[0.0, 1.0]$. We assume that all the agents in a team can share their observations, therefore the set of wordline components that are observable by a team τ, S_τ is the union of the sets of all of its members:

$$S_\tau = \bigcup_{i \in \tau} S_i . \qquad (7)$$

The percent of agents in other teams that can be observed by a team τ of size m is therefore:

$$B_\tau = 1 - (1 - B_i)^m , \qquad (8)$$

assuming that the set of agents in other teams that each team member can observe sampled independently for all team members. Any costs associated with team members sharing information can be included in the global utility. In Section IV we will examine issues in domains where there is a high cost of sharing observations within a team.

III. THE BAR PROBLEM

Arthur's bar problem [1] can be viewed as a problem in designing collectives. Loosely speaking, in this problem at each time step each agent decides whether to attend a bar by predicting, based on its previous experience, whether the bar will be too crowded to be "rewarding" at that time, as quantified by a utility function G. The selfish nature of the agents frustrates the global goal of maximizing G. This is because if most agents think the attendance will be low (and therefore choose to attend), the attendance will actually be high, and vice-versa.

Here, we focus on the following more general variant of the bar problem investigated in [20]: There are N agents broken up into disjoint teams, where each agent goes to the bar one night each week. The action of the agent is to choose the one

night (out of seven) it will go to the bar that week. At the end of the week, each agent receives the team utility for its team. The task of the agent is to choose a night that maximizes its team utility.

More formally, the global utility in any particular week is:

$$G(z) \equiv \sum_{k=1}^{7} x_k(z) \exp(-x_k(z)/c) ,\qquad (9)$$

where $x_k(z)$ is the total attendance on night k and c is a real-valued parameter. In this problem when either too few or too many agents attend some night in some week the global utility G is low.

Since we wish to concentrate on the effects of the utilities rather than on the RL algorithms that use them, we use a (very) simple RL algorithm. In our algorithm each agent τ has a 7-dimensional vector giving its estimates of the team utility it would receive for choosing each possible night. The decisions are made using the vector, with an ϵ-greedy learner with ϵ set to 0.05. All of the vectors are initially set to zero and there is a learning rate decay is 0.99. The RL algorithm can be viewed as a Q-learner with $\gamma = 0$. Note that many other utility maximization algorithms could be used instead, including evolution over neural networks.

This paper uses a version of this problem where teams cannot fully observe the actions of agents in other teams. The number of actions that can be observed is dependent on the team size, and is based on collecting the observations of all the members in a team. Let the observation level B, be the fraction of agents from other teams that can be observed by a single agent. The range of B is $[0.0, 1.0]$. A team is able to aggregate all the observations collected by its team members. This aggregation will make the observation level of the team B_τ significantly higher than the observation level of any particular agent, and will rise with the number of the agents in the team.

In the Bar Problem, partial observability influences how $x_k(z)$ is computed. For truncated versions of the DU, (BTU and TTU), we use $x_k(z^{o_\tau})$ which returns how many of the observable patrons are going on night k (note since in BTU the first term is broadcast, the team does not need to compute it). For utilities using an estimate of the state (BEU and EEU), $x_k(z^o)$ is scaled, and $\frac{1}{B_\tau} x_k(z^{o_\tau})$ represents the estimate of how many agents actually went on night k. For example when $B_\tau = 0.25$, we assume that $x_k(z^{o_\tau})$ is really only accounting for one quarter of the agents, so we scale it by $\frac{1}{0.25} = 4$. Note this is an extremely simple estimation procedure and does not take any information an agent collects to modify how it forms this estimate.

IV. RESULTS

We tested the performance of each of the four version of the DU in Arthur's Bar Problem. Each team utility was tested with a combination of fourteen different team sizes and eleven levels of observability for a total of 154 tests per utility. Each test was conducted with 100 agents and with c equal to five.

All of the trials were conducted for 1000 episodes, and were run 25 times. These tests measured the relative merits of the four team utilities as well as possible benefits of changing team sizes. The results show that in some partially observable domains, changing the utility can increase the performance of the system, but that changing the team size without changing the utility may be the best way to increase performance.

A. Domain without Information Sharing Costs

Figure 2 shows the tradeoffs between choices of team size at different levels of agent observability in a domain where there is no costs associated with agents sharing observations with other team members. The observational capabilities of the team at each point can be inferred from equation 8. These results show that the BEU and TTU team utilities are usually worse than BTU or EEU. In addition teams using EEU almost always perform better than teams using BTU. However, Figure 3 shows that EEU has difficulties with small team sizes and low levels of observability. When an agent in a team can observe only 10% of agents in other teams, BTU is the best utility when each team only has one agent. Team utility BTU is superior in this low observability case, since it uses the broadcast of the global utility to overcome the lack of observations. In this case, a system designer using EEU may consider using BTU instead. However, expanding the team size while continuing to use the EEU is an even better option. Even when each team only has two agents, the combined observational capability of these agents enables the teams to do far better when they use the BTU. When teams of three are formed, the system using EEU performs 50% better than the system using the next best utility, BTU. The systems using small teams with the EEU even performs 30% better than a single team system (team size = 100) with full observability. This happens because agents have difficulty maximizing a team utility for a single team system, since the utility is influenced by the actions of all of the agents in the entire system. In general when teams are too large, the performance of the system goes down, even when EEU is used. The best team size is typically around five or ten agents. This optimum represents the best balance between having small teams with more learnable team utilities and large teams, which collect more observations.

With the non-factored utilities, EEU and TTU, this balance of team size comes from the tradeoff between factoredness and learnability. Even though as team sizes get smaller, they become more learnable, they also become less factored since as information sharing goes down causing the first term in the difference equation to diverge from G. For the factored utilities BEU and BTU, there is a tradeoff between two different ways noise comes into the system. When teams are large, more components are removed from the second term of the difference equation allowing more noise from the first term to remain. When teams are small, the lack of information sharing has a similar effect, in that many of the components in the second term are not included because their values are unknown.
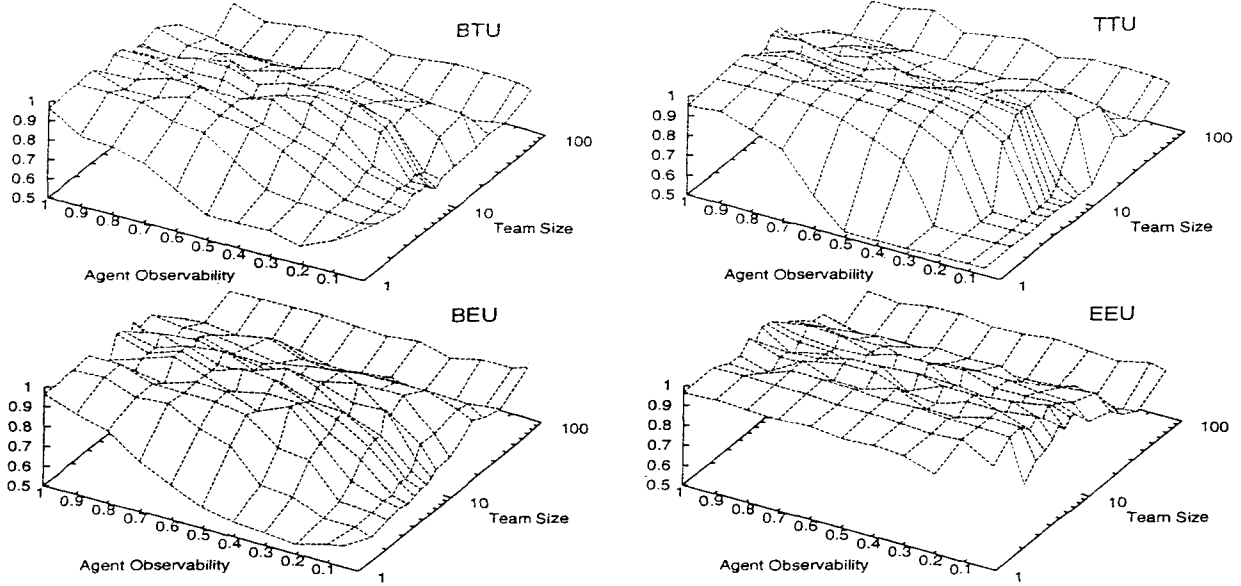
Fig. 2. Performance with different team sizes and observability. Each graph is for a different utility. From top-left, clockwise the utilities used are: BTU, TTU, EEU, BEU. The two utilities, BEU and EEU, that estimate hidden values rather than ignoring them, perform much better than their conterparts, BTU and TTU.
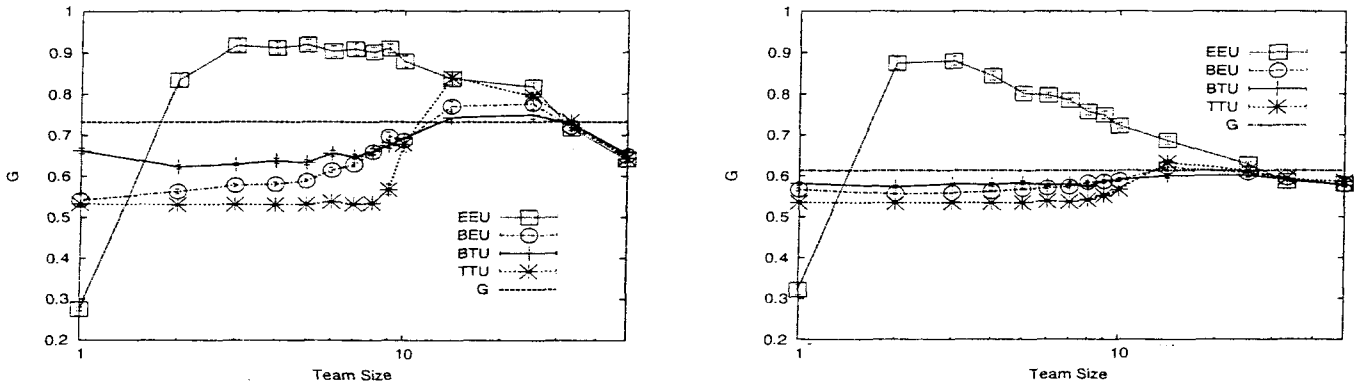

Fig. 3. Performance of four utility functions at 10% observability. EEU performs best for most team sizes under normal learning time (left).The signal to noise advantages of EEU become more apparent when learning time is reduced to 1/8 of original time (right).

B. Domains with Information Sharing Costs

The global utility used for previous experiments does not include any possible communication cost associated with team members sharing their observations with the rest of the team. This global utility models situations for tightly bound teams, where it is easy to share information in addition to domains where an existing team-sharing infrastructure already exists. Since there are no team-sharing costs, it is assumed that team members always share their observations with the rest of the team. However, in some domains there may be a significant cost to share observations with other team members. This section will explore whether it is beneficial in such a domain for the agents to be able to choose whether or not to share their observations with the rest of the team.

Assuming that teams are using the EEU as their team utility, we can incorporate sharing costs by subtracting it off the utility. The EEU including sharing costs, EEU^C, is

defined as:

$$EEU_\tau^c(z) = EEU_\tau(z) - m(z)C , \qquad (10)$$

where C is the cost for one team member to share its observations and $m(z)$ returns the number of team members that choose to share. In addition to incorporating the sharing costs into the utility, a binary action is added to each agent's actions space: the choice of whether or not to share observations with the rest of the team. Note that this addition doubles the number of possible action choices an agent can make and could make it significantly harder for agents to learn. In addition, this choice of action could make the team utility more variable since it changes continuously as team members decide whether or not to share.

To test the effectiveness of allowing agents to choose whether or not to share information, we performed experiments where the cost of information sharing was $C = \frac{m\hat{G}}{100n}$ where there are m agents in each group out of a total of
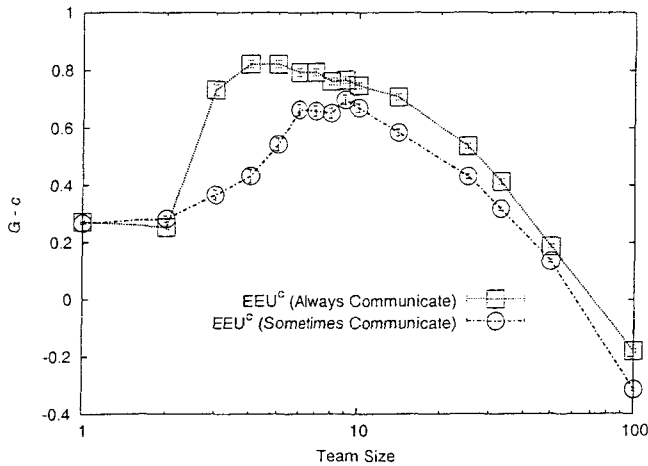
Fig. 4. When agents can choose whether or not to share information, the problem becomes much more difficult and performance goes down.

n agents, and $\hat{G}$ is the maximum possible value of G.[2] Figure 4 shows that agents that could choose whether to share information performed worse than agents that always shared information. Agents using EEU^c that always shared information performed the best, since there was much less noise in this scenario. When agents could choose to share, the utility was less stable since it could change dramatically depending on whether other agents in the team chose to share at a particular time step. Note that adding sharing cost did not significantly change the dynamics of the system. Team with five members still performed the best. However, the performance drop-off with large teams was even greater since here large teams have more sharing costs in addition to having less learnable team utilities.

V. DISCUSSION

In this work we focus on the problem of designing a collective of teams of autonomous agents in the presence of partial observability. In such cases, team utilities which rely on teams fully observing other teams may break down. We presented four different utility functions that each make different tradeoffs among what information is available to a team and how that information should be used.

We saw that there are tradeoffs involved in choosing team sizes. Large teams can benefit from increased observation capability. However the team utilities of small teams are easier to learn. We showed that team sizes can be adjusted to make the optimal balance between these tradeoffs.

Furthermore, in many problems agents can choose whether to share information or not, and consequently incur a cost or not. Preliminary results show that in such cases, agents have a difficult time in learning to maximize the team utility function. This difficulty is due to the constant change in the value of the team utility, which depends on the sharing choices of many other agents. This constant change in effect creates a more noisy learning environment.

REFERENCES

[1] W. B. Arthur. Complexity in economic theory: Inductive reasoning and bounded rationality. *The American Economic Review*, 84(2):406–411, May 1994.
[2] R. Axtell. Non-cooperative dynamics of multi-agent teams. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1082–1089, Bologna, Italy, July 2002.
[3] P. Busetta, A. Dona, and M. Nori. Channeled multicast for group communications. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1280–1287, Bologna, Italy, July 2002.
[4] P. Cohen and H. Nous Levesque. Teamwork. In *Special Issue on Cognitive Science and AI, 25, 4*, pages 487–512. 1991.
[5] R. H. Crites and A. G. Barto. Improving elevator performance using reinforcement learning. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems - 8*, pages 1017–1023. MIT Press, 1996.
[6] D. Fox, W. Burgard, H. Kruppa, and S. Thrun. A probabilistic approach to collaborative multi-robot localization. *Autonomous Robots*, 2000.
[7] D. Gage. How to communicate with zillions of robots. In *Proceedings of SPIE Mobile Robots V III*, pages 250–257, Boston, MA, 1993.
[8] A. Greenwald, E. Friedman, and S. Shenker. Learning in network contexts: Experimental results from simulations. *Journal of Games and Economic Behavior: Special Issue on Economics and Artificial Intelligence*, 35(1/2):80–123, 2001.
[9] B. Grosz and S. Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86:269–358, 1996.
[10] Sandip Sen, Mahendra Sekaran, and John Hale. Learning to coordinate without sharing information. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 426–431, Seattle, WA, 1994.
[11] P. Stone and M. Veloso. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots*, 8(3), 2000.
[12] Ashley Stroupe, Martin C. Martin, and Tucker Balch. Distributed sensor fusion for object position estimation by multi-robot systems. In *IEEE International Conference on Robotics and Automation*. IEEE, May 2001.
[13] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.
[14] S. Talukdar, L. Baerentzen, A. G., and P. de Souza. Asynchronous teams: Cooperation schemes for autonomous agents. *Journal of Heuristics*, pages 295–321, Dec 1998.
[15] K. Tumer and A. Agogino. Overcoming communication restrictions in collectives. July 2004. Submitted to IJCNN.
[16] K. Tumer, A. Agogino, and D. Wolpert. Learning sequences of actions in collectives of autonomous agents. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 378–385, Bologna, Italy, July 2002.
[17] K. Tumer and D. H. Wolpert. Collective intelligence and Braess' paradox. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 104–109, Austin, TX, 2000.
[18] M. P. Wellman. A market-oriented programming environment and its application to distributed multicommodity flow problems. In *Journal of Artificial Intelligence Research*, 1993.
[19] D. H. Wolpert, S. Kirshner, C. J. Merz, and K. Tumer. Adaptivity in agent-based routing for data networks. In *Proceedings of the fourth International Conference of Autonomous Agents*, pages 396–403, 2000.
[20] D. H. Wolpert and K. Tumer. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, 4(2/3):265–279, 2001.
[21] D. H. Wolpert, K. Tumer, and J. Frank. Using collective intelligence to route internet traffic. In *Advances in Neural Information Processing Systems - 11*, pages 952–958. MIT Press, 1999.
[22] D. H. Wolpert, K. Wheeler, and K. Tumer. Collective intelligence for control of distributed dynamical systems. *Europhysics Letters*, 49(6), March 2000.

---

[2] The maximum average utility per agent is $\hat{G}/n$. With a 1% sharing cost per each agent, the cost is $\hat{G}/100n$. When sharing with m group members the cost is $m\hat{G}/100n$.